Grenze International Journal of Engineering and Technology, June Issue

Reliability-Aware CNN–SNN-Inspired Fusion with Adaptive Modality Dominance for Road-user Classification

P. Madhavi[1], B. Krishna Gopala Swami[2], K. Manikanta Reddy[3] and Y. Abhi Sai Reddy[4]
[1-4]Dept. of Artificial Intelligence and Data Science, Lakireddy Bali Reddy College of Engineering (Autonomous), Mylavaram, India
Email: pmadhavi@lbrce.ac.in, krishnagopalswami183@gmail.com, manikantareddy87121@gmail.com, abhireddy8484@gmail.com

***Abstract*— Vision-based traffic participant classification systems often rely on a single visual cue, making them vulnerable to real-world degradations such as motion blur, occlusion, or poor illumination. Spatial models such as convolutional neural networks (CNNs) excel under clear conditions but degrade with image corruption, while temporal models such as spiking neural networks (SNNs) capture motion dynamics but fail under temporal disruption. We propose Adaptive CNN–SNN Fusion (ACSF), a framework that dynamically balances spatial CNN and temporal SNN-inspired experts on a per-input basis through a learnable fusion gate. Unlike static fusion, ACSF learns input-dependent weights reflecting each modality's reliability. Experiments on custom dashcam and BDD100K datasets show that the fusion gate adapts its weighting under controlled degradations, maintaining accuracy within 1% drop while outperforming single-modality and fixed-fusion baselines, confirming adaptive expert fusion's effectiveness for robust traffic perception.**

***Index Terms*— Adaptive fusion, CNN, SNN, Traffic participant classification, Dashcam vision, Robust perception.**

I. INTRODUCTION

Autonomous driving and intelligent transportation systems require reliable perception of traffic participants such as vehicles, pedestrians, and cyclists. Vision-based classification approaches typically depend on spatial appearance cues extracted from individual frames or temporal motion cues derived from video sequences. While recent deep learning models achieve high accuracy under controlled conditions, their performance often degrades in real-world scenarios characterized by motion blur, occlusion, sensor noise, and variable illumination. CNNs excel at extracting spatial features from high-quality images but are sensitive to visual corruption. Under strong blur, low light, or partial occlusion, spatial evidence can become ambiguous. In contrast, SNNs capture motion dynamics and temporal consistency by integrating information across multiple frames [1], [2], [3]. These models can remain confident when individual frames are degraded, but may fail when frame order or temporal continuity is disrupted, for example due to dropped frames or irregular sampling.
Existing fusion approaches often combine spatial and temporal cues using static weights or simple late fusion strategies [4], [5].

Grenze ID: 01.GIJET.12.2.506_1

Such designs implicitly assume that both modalities are equally reliable across all inputs. In real-world driving, this assumption rarely holds. A night-time scene with clear motion but poor contrast favors temporal reasoning, whereas a single sharp snapshot under good illumination may suffice for a purely spatial model. Treating both experts as uniformly trustworthy ignores input-dependent variability and can lead to suboptimal behavior.

To address this limitation, we propose ACSF that dynamically adjusts the contribution of spatial and temporal experts on a per-input basis through a learnable fusion gate. Unlike prior work targeting energy efficiency or neuromorphic deployment of SNNs [6], [7], our goal is to study reliability-aware fusion behavior in a standard deep-learning setting. The main contributions are: (i) ACSF coupling a spatial CNN expert with a temporal SNN-inspired expert via a learnable fusion gate, enabling per-clip adaptive weighting; (ii) controlled spatial and temporal degradation protocols to probe reliability-aware fusion behavior without retraining; and (iii) detailed empirical analysis showing how the fusion gate reallocates trust between experts under clean, CNN-degraded, and SNN-degraded conditions.

II. Related Work

A. Spiking Neural Networks and Hybrid Models

SNNs have emerged as a promising paradigm for event-based and temporal data due to their sparse, event-driven computation and biological plausibility [3]. Martinez et al. [1] investigated SNN-based object detection and reported competitive performance and energy advantages. Miramond and Thierion [2] demonstrated effective processing of automotive event-camera data. Zhang et al. [5] introduced a hybrid spiking fully convolutional network for semantic segmentation. ReSpike [8] and Spike-HAR++ [9] applied residual and transformer-style spiking architectures to action recognition. DT-SCNN [10] further optimized spike-based convolutions for edge devices. Our work leverages a temporal SNN-inspired expert with a CNN, but operates on RGB dashcam clips using rate-based approximations, focusing on fusion behavior analysis rather than hardware deployment.

B. Event-based Vision and Multi-Modal Fusion

Shariff et al. [11] reviewed event cameras for automotive sensing, highlighting their low latency and high dynamic range. Mimouna [4] explored deep-learning-based data fusion for multi-object detection. Sun [12] demonstrated that fusing event-based and frame-based vision improves moving-object detection. Steffen et al. [13] proposed a gesture recognition system fusing event-based and depth data within a spiking CNN. Pan et al. [14] proposed multi-scale evolutionary search for deep spiking networks. Wang et al. [6] presented a storage-efficient SNN–CNN hybrid for traffic sign recognition. Unlike these efforts optimizing SNN architectures, we study how a learned fusion gate can modulate expert influence in response to changing input reliability.

C. Positioning of This Work

Most prior work on CNN–SNN hybrids emphasizes performance, efficiency, or hardware deployment. Far fewer studies investigate how fusion decisions evolve when different modalities are selectively degraded. Table I compares ACSF with representative approaches. The key distinction of ACSF is its learnable fusion gate that dynamically allocates trust between spatial and temporal experts based on input quality, enabling reliability-aware behavior without explicit supervision.

TABLE I: COMPARISON WITH EXISTING CNN-SNN HYBRID METHODS

Study	Method	Dataset	Key Limitation
Martinez et al. [1]	Fully spiking SNN detector	Automotive object detection	No CNN-SNN fusion or per-sample gate
Miramond et al. [2]	Event-camera SNNs	Event-based driving data	No RGB clips or fusion gate
Mimouna [4]	Multi-modal fusion for ITS	OLIMP traffic data	No learnable reliability gate
Zhang et al. [5]	Hybrid spiking FCN	VOC2012, event-based	Segmentation, not classification
Xiao et al. [8]	Residual hybrid SNN	Video action recognition	No reliability-based fusion
Wang et al. [6]	RRAM-based SNN-CNN	Traffic sign recognition	Hardware-focused, no degradation study
ACSF (ours)	Adaptive CNN-SNN fusion gate	Dashcam + BDD100K	Analyzes reliability-aware fusion under degradations

III. Dataset Collection and Preparation

A. Source and Acquisition

We constructed a custom dataset from publicly available dashcam videos sourced from YouTube, covering a range of traffic densities, road geometries, weather conditions, and camera viewpoints. Raw videos were

manually curated to exclude unrealistic footage. Each video was sampled at 5 fps, and we extracted short clips of six consecutive frames (∼1.2 seconds), providing enough temporal context for motion consistency without excessive computational cost.

B. Classes, Annotation, and Preprocessing

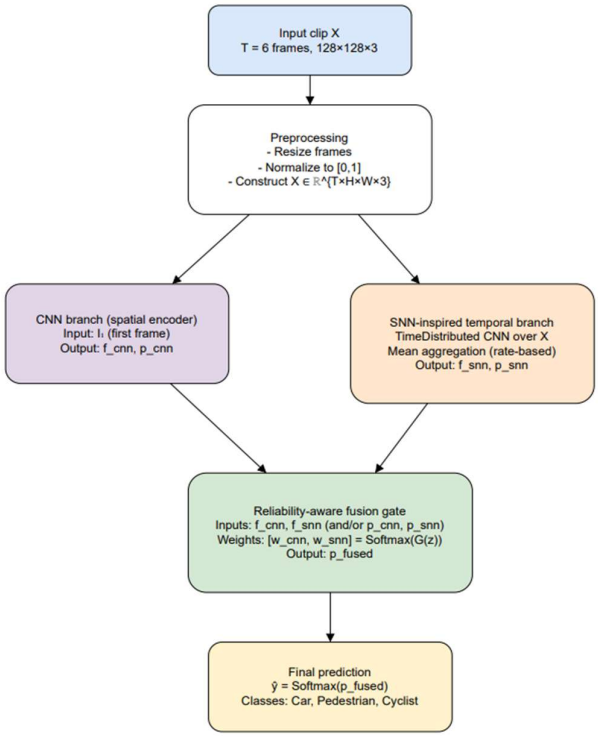
We focus on three road-user categories: Car (motor vehicles including sedans, SUVs, and small vans), Pedestrian (single walking or standing persons), and Cyclist (bicyclists or two-wheeled non-motorized users). Annotations were performed in multiple passes with verification by a second annotator. To avoid temporal leakage, we partitioned the dataset at video level. All frames were resized to 128×128 pixels using bilinear interpolation and normalized to [0, 1]. The CNN branch uses only the first frame; the temporal branch receives the full six-frame sequence. During training, we apply mild augmentation (random horizontal flips and small brightness jitter). Class-weighted losses mitigate natural imbalance.

C. BDD100K Benchmark Dataset

To validate generalizability, we evaluate on BDD100K [15], which provides large-scale diverse driving sequences across various weather conditions, times of day, and geographic locations. We extract six-frame clips labeled as Car, Pedestrian, or Cyclist using official splits. The validation set contains 1,910 clips (1,694 Car, 203 Pedestrian, 13 Cyclist), reflecting real-world traffic distributions. All preprocessing steps (resizing, normalization) remain identical to the custom dataset.

IV. Model Architecture

The proposed ACSF framework consists of three components: a spatial CNN expert, a temporal SNN-inspired expert, and an adaptive fusion gate.



A. Spatial CNN Expert

The CNN branch processes a single representative frame from each clip. It consists of three convolutional blocks, each containing a 3×3 convolution, batch normalization, ReLU activation, and 2×2 max pooling. The final feature map is followed by global average pooling and a fully connected layer that maps to a d-dimensional embedding f_CNN (d = 128), producing softmax output p_CNN ∈ R^3 over three classes.

B. Temporal SNN-Inspired Expert

The temporal expert operates on the full six-frame sequence inspired by rate-based SNN formulations [3], [17]. Each frame I_t is passed through a shared-weight TimeDistributed CNN: $h_t = F_{TD}(I_t)$. Frame-wise embeddings are aggregated by mean: $f_{SNN} = (1/T)\Sigma h_t$, approximating the rate-coding principle where firing rate encodes stimulus intensity. The temporal expert has a similar structure to the CNN branch but with reduced channel widths to control model size.

C. Adaptive Fusion Gate

High-level feature embeddings are concatenated: $z = [f_{CNN}; f_{SNN}] \in R^{2d}$. This descriptor is passed through a lightweight fully connected network (hidden layer with ReLU, followed by a linear layer with two outputs). A softmax layer produces: $[w_{CNN}, w_{SNN}] = \text{Softmax}(W_g z + b_g)$, ensuring $w_{CNN} + w_{SNN} = 1$. The final prediction is: $p_{fused} = w_{CNN} p_{CNN} + w_{SNN} p_{SNN}$. To prevent degenerate solutions where one expert always dominates, we introduce entropy-based regularization: $L_{ent} = -\lambda \Sigma_m w_m \log w_m$.

V. Optimization and Training

ACSF is trained end-to-end using a multi-objective loss: $L = L_{fused} + \lambda_{CNN} L_{CNN} + \lambda_{SNN} L_{SNN} + L_{ent}$, where each cross-entropy term provides gradient signal to the corresponding branch. λ_{CNN} and λ_{SNN} are set to 0.3. Class weights are applied inside cross-entropy terms to mitigate natural imbalance between Car, Pedestrian, and Cyclist classes. We optimize using Adam (learning rate 10^{-3}, mini-batches of 32 clips) for up to 30 epochs with early stopping. Dropout (rate 0.2) is included in both experts and fusion gate. In a second phase, we optionally freeze expert weights and fine-tune only the fusion gate, allowing rapid adaptation to new conditions without full retraining.

VI. Experimental Results

A. Baseline Performance

Table II reports classification accuracy on clean test data for single-expert baselines and ACSF. ACSF outperforms both single-expert baselines on both datasets, confirming that combining spatial and temporal information yields measurable gains even under nominal conditions.

Table II: Classification Accuracy on Clean Test Sets

Model	Dashcam (%)	BDD100K (%)
CNN only (spatial)	96.7	91.8
SNN only (temporal)	94.1	92.0
ACSF (proposed)	98.1	92.8

B. Fusion Baseline Comparison

Table III compares ACSF against several fusion strategies on BDD100K. Early Fusion (92.3%) underperforms methods preserving separate expert predictions. Late Fusion and Fixed Fusion both achieve 92.8%, matching ACSF on clean data.

ACSF's advantage becomes apparent under degraded conditions, where adaptive weight modulation maintains higher accuracy than fixed strategies.

Table III: Comparison of Fusion Strategies on BDD100K (CLEAN)

Fusion Method	Accuracy (%)
CNN-only (no fusion)	91.8
SNN-only (no fusion)	92.0
Early Fusion	92.3
Late Fusion (average)	92.8
Fixed Fusion (0.5/0.5)	92.8
ACSF (adaptive)	92.8

C. Robustness Under Controlled Degradation

Tables IV and V report accuracy under controlled degradations (Gaussian blur, darkening, contrast reduction, additive noise) applied at test time without retraining. On BDD100K, standalone CNN accuracy drops to 89.5% under spatial degradation, but ACSF maintains 92.0% by leveraging the intact SNN branch. When SNN is

degraded (89.7% standalone), ACSF achieves 92.3% by relying more on the CNN. Performance drops are minimal (0.8% and 0.5% respectively), demonstrating effective compensation.

TABLE IV: TEST ACCURACY UNDER CONTROLLED DEGRADATION (CUSTOM DASHCAM)

Condition	Accuracy (%)
Clean	98.1
CNN-degraded	95.1
SNN-degraded	96.5

TABLE V: TEST ACCURACY UNDER CONTROLLED DEGRADATION (BDD100K)

Condition	Fused (%)	CNN (%)	SNN (%)
Clean	92.8	91.8	92.0
CNN-degraded	92.0	89.5	92.0
SNN-degraded	92.3	91.8	89.7

D. Fusion Adaptability Analysis

Tables VI and VII report mean fusion weights under different conditions. Under clean conditions, the gate favors the CNN expert (0.93 on Dashcam, 0.72 on BDD100K). Under CNN-degraded conditions, w_CNN drops dramatically (0.59/0.18) and w_SNN increases correspondingly. When temporal input is corrupted, the gate shifts weight to the CNN (0.96/0.82). The larger weight shift on BDD100K suggests that its greater visual diversity makes the gate more sensitive to modality reliability changes, precisely the intended functionality of ACSF.

TABLE VI: MEAN FUSION WEIGHTS UNDER DIFFERENT CONDITIONS (CUSTOM DASHCAM)

Condition	w_CNN	w_SNN
Clean	0.93	0.07
CNN-degraded	0.59	0.41
SNN-degraded	0.96	0.04

TABLE VII: MEAN FUSION WEIGHTS UNDER DIFFERENT CONDITIONS (BDD100K)

Condition	w_CNN	w_SNN
Clean	0.72	0.28
CNN-degraded	0.18	0.82
SNN-degraded	0.82	0.18

VII. DISCUSSION

The experiments demonstrate that the ACSF fusion gate responds meaningfully to input conditions rather than collapsing to a single expert. On clean data, the gate naturally aligns with the strongest expert (CNN), while under spatial degradations it compensates by increasing reliance on the temporal SNN-inspired branch, and reverses this behavior under temporal degradations. This confirms our core hypothesis that expert reliability varies across scenarios and can be learned directly from data.

On clean data, adaptive fusion performs equivalently to simpler late-fusion or fixed-weight strategies. However, under degradation, Fixed Fusion (w_CNN = w_SNN = 0.5) continues to give equal weight to both experts even when one is severely corrupted, whereas ACSF dynamically reallocates trust—shifting w_SNN from 0.28 to 0.82 under CNN degradation.

Importantly, ACSF achieves this without explicit reliability labels or hand-crafted fusion rules, suggesting similar mechanisms could be applied to other complementary modalities such as event-based and frame-based vision [11], [12] or radar and camera data [18].

Limitations include: short temporal clips (six frames) may not capture complex behaviors; rate-based aggregation differs from true neuromorphic evaluation; severe class imbalance in BDD100K (only 13 Cyclist samples vs. 1,694 Car) results in zero recall for the minority class; and the custom dashcam dataset may not fully represent production automotive sensor characteristics.

Future work will extend ACSF to longer sequences, additional road-user classes, true multi-sensor setups (radar+camera), architecture search [14] to jointly optimize experts and gate, and deployment on neuromorphic hardware [6], [7].

VIII. Conclusion

This paper presented ACSF, an adaptive fusion framework that dynamically balances spatial CNN and temporal SNN-inspired experts for robust traffic participant classification. Unlike static fusion approaches, ACSF learns input-dependent weights reflecting the relative reliability of each modality through a lightweight fusion gate trained end-to-end. Experiments on custom dashcam and BDD100K datasets demonstrate that the fusion gate automatically shifts weight between experts when one modality is degraded, maintaining accuracy with minimal performance drop (less than 1%). The proposed approach outperforms single-modality baselines and early fusion, matches late fusion on clean data, and surpasses fixed-weight fusion under degraded conditions. Visualization of fusion weights confirms interpretable reliability-aware behavior without explicit supervision.

References

[1] S. A. Martinez, D. Ser, J. Garcia-Bringas, and P. Sabbella, "Efficient object detection in autonomous driving using spiking neural networks," in Proc. IEEE 26th Int. Conf. on Intelligent Transportation Systems (ITSC), IEEE, 2023, pp. 1–8.

[2] L. C. B. Miramond and P. Thierion, "Object detection with spiking neural networks on automotive event data," arXiv preprint, 2022.

[3] H. Sabbella et al., "The promise of spiking neural networks for ubiquitous computing: A survey and new perspectives," arXiv preprint arXiv:2506.01737, 2025.

[4] A. Mimouna, "Exploring data fusion for multi-object detection for intelligent transportation systems using deep learning," Ph.D. dissertation, Université Polytechnique Hauts-de-France and Université de Sousse, 2021.

[5] T. Zhang, S. Xiang, W. Liu, Y. Han, X. Guo, and Y. Hao, "Hybrid spiking fully convolutional neural network for semantic segmentation," Electronics, vol. 12, no. 17, p. 3565, 2023.

[6] Z. Wang, F. A. Ghaleb, A. Zainal, M. M. Siraj, and X. Lu, "A storage-efficient SNN–CNN hybrid network with RRAM-implemented weights for traffic signs recognition," Scientific Reports, 2024.

[7] A. L. L. Oliveira, M. A. R. de Oliveira, and J. P. Papa, "Event-based spiking neural networks for visual object recognition: A review," IEEE Access, vol. 12, pp. 51275–51306, 2024.

[8] L. Xiao, Y. Li, Y. Kim, D. Lee, and P. Panda, "ReSpike: Residual frames-based hybrid spiking neural networks for efficient action recognition," arXiv preprint, 2024.

[9] X. Lin, M. Liu, and H. Chen, "Spike-HAR++: An energy-efficient and lightweight parallel spiking transformer for event-based human action recognition," Frontiers in Computational Neuroscience, vol. 18, p. 1508297, 2024.

[10] F. Lei, X. Yang, J. Liu, R. Dou, and N. Wu, "DT-SCNN: Dual-threshold spiking convolutional neural network for edge applications," Frontiers in Computational Neuroscience, vol. 18, p. 1418115, 2024.

[11] W. Shariff et al., "Event cameras in automotive sensing: A review," IEEE Access, vol. 12, pp. 51275–51306, 2024.

[12] H. Sun, "Moving objects detection and tracking using hybrid event-based and frame-based vision for autonomous driving," Ph.D. dissertation, École Centrale de Nantes, 2023.

[13] L. Steffen et al., "Efficient gesture recognition on spiking convolutional networks through sensor fusion of event-based and depth data," in Proc. IEEE Int. Conf. on Robotics and Automation (ICRA), IEEE, 2024.

[14] W. Pan et al., "Multi-scale evolutionary neural architecture search for deep spiking neural networks," arXiv preprint arXiv:2304.10749, 2023.

[15] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2020, pp. 2636–2645.

[16] A. G. Gonzalez and A. S. Kumar, "Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," Array, vol. 10, p. 100057, 2021.

[17] Q. Su, W. He, X. Wei, B. Xu, and G. Li, "Multi-scale full spike pattern for semantic segmentation," Neural Networks, vol. 176, p. 106330, 2024.

[18] A. Shaaban, "Spiking neural networks for event-driven FMCW radar signal classification," Ph.D. dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2025.